O10

ディープラーニングに基づく化合物とタンパク質の相互作用予測

(京大院医[1]，理研[2]，先端医療振興財団[3])
○浜中雅俊[1]*，種石　慶[2]，岩田浩明[3]，奥野恭史[1]

1. はじめに

本稿では，医薬品となる化合物を発見するための第一段階のスクリーニングであるタンパク質と化合物の相互作用の予測について，多層のニューラルネットワークであるディープラーニング（深層学習）を用いた試みについて述べる．

膨大な種類の化合物から医薬品になり得るリガンド化合物を見つけ出す工程は，開発にかかる時間とコストを押し上げる主要因となっており，計算により優れた性質の候補化合物を絞り込むインシリコ（in silico）スクリーニングの手法が提案されてきた．

ドッキングシュミレーション法[1-5]では，化合物とタンパク質の立体構造を用いて分子の結合を予測することを可能としていたが，長時間の計算が必要なことや立体構造が未知の化合物では予測が困難であるという問題があった．

我々は，相互作用が確認された 12.5 万件の結合データと，結合データに含まれない同数の組み合わせを非結合データとして用意し，それらをサポートベクターマシン（SVM）で学習することで新たなデータに対して相互作用を予測する CGBVS 法（Chemical Genomics-Based Virtual Screening）を提案してきた[6]． CGBVS 法では，学習するデータが増えるにつれて SVM による計算時間や計算機上のメモリ使用量が増大するため，100 万件を超える大規模な相互作用データを用いて現実的な時間で学習することは困難であった．

そこで我々は，100 万件を超える大規模な相互作用データの学習を現実的な時間および計算量で行うことを目指した手法として，ディープラーニングの一手法である DBN（Deep Belief Networks）[7,8]に基づく予測手法，CGBVS-DBN 法を提案し，1 万件の相互作用データを用いて性能の検証を行ってきた[9]．その結果，データ件数が多いほど性能が向上すること，中間層数やユニット数によって性能が変化することを確認した．

本稿では，25 万件の相互作用データを用いて，CGBVS-DBN 法と従来の CGBVS 法との予測性能の比較を行う．

2. 相互作用データ

5 節の実験で用いる 25 万件の相互作用データは，相互作用があることを実験により確認している正例 12.5 万件と，人工的に生成した負例 12.5 万件からなる[10].

正例 12.5 万件の抽出

正例データの抽出は，まず，複数の相互作用データベースの情報を統合し，重複を排除した 200 万件のタンパクと化合物の組の中からランダムで 12.5 万件を抽出する．次に，タンパク質については，PROFEAT[11]を用いてアミノ酸配列を 1080 次元のベクトルで表現し，化合物については，Dragon[12]を用いて化学構造を 894 次元のベクトルで表現する．この合計 1974 次元のベクトルで表現したものを正例とする．

負例 12.5 万件の生成

市販あるいは公開されている相互作用データベースにおけるデータの大部分は相互作用「あり」の化合物とタンパク質の組み合わせであり，相互作用「なし」の組み合わせはわずかである．

そこで本研究では，以下のアルゴリズムにより人工的に負例を発生させている．まず，12.5 万件の順序をランダムにシャッフルする． 次に，順序が上の負例から順に交差負例を生成していく．具体的には，まず，i 番目と i+1 番目の正例の化合物とタンパクの組み合わせを入れ替える．すなわち，i 番目の化合物と i+1 番目のタンパク質の組と，i+1 番目の化合物と i 番目のタンパク質の組を作成して，それらを負例候補とする．そして，負例候補が正例に含まれていない場合には，確定負例とし，次の２つの正例（i+2 番目と i+3 番目の正例）の組み換えを行う．負例候補が正例に含まれる場合には，負例候補と次の正例との組み換えを行う．このような交差負例の発生を順に行い，12.5 万件すべてを確定負例とする．ただし，最後に正例が 1 つだけ残った場合や，正例と重複した負例候補のみが残された場合には，それを確定負例にすることできないため取り除く．また，生成した確定負例間での重複についても取り除く．

*masatosh@kuhp.kyoto-u.ac.jp

3. CGBVS 法

本稿で提案する CGBVS-DBN 法は，従来提案されていた CGBVS 法における機械学習の部分を SVM から DBN に変更したもので，両者の予測の枠組み自体は共通している．

図 1 は，CGBVS 法の流れを表したものである．まず，学習データである化合物とタンパク質の相互作用データについて，Dragon および PROFEAT を用いて記述子で表現することで，1974 次元の入力ベクトルを用意する．このとき，化合物とタンパク質の組み合わせには，正例（相互作用あり）もしくは負例（相互作用なし）のラベルが与えられているため，正例を 1，負例を 0 とした 1 次元の出力ベクトルとする．

次に，SVM を用いて入力ベクトルと出力ベクトルとの関係を学習する．学習により獲得したモデルを用いることで，新たな化合物とタンパク質の組に対して相互作用があるか，あるいは，ないかを予測することが可能となる．

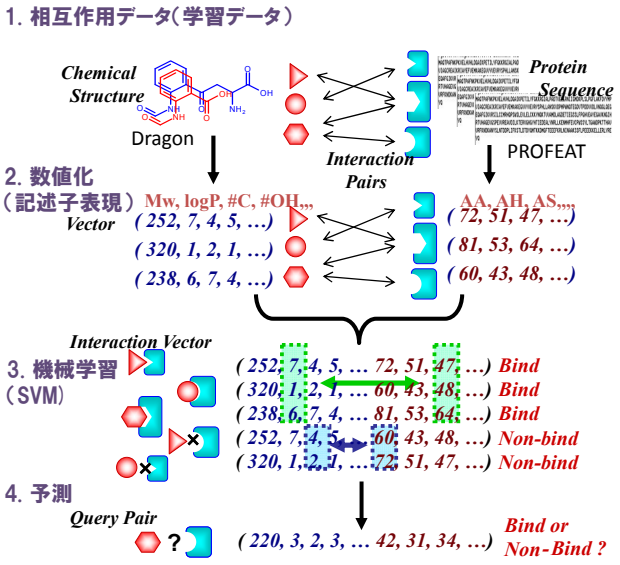

図 1：CGBVS 法

4. CGBVS-DBN 法

CGBVS 法は SVM を用いて機械学習をしているのに対し，CGBVS-DBN 法は DBN を用いて学習を行う． DBN は，プレトレーニングと呼ばれる教師なし学習を行うことで，これまで困難であった多層のニューラルネットワークの学習を可能にしたものである．

図 2 は，我々が用いた DBN の構成を示したもので，入力はタンパク質および化合物の記述子から得た 1974 次元の特徴を平均 0 分散 1 で正規化した値，出力は相互作用がある場合に 1，ない場合に 0 を示す 2 ユニットである．入力層と 1 段目の中間層および中間層とその次の中間層の間のネットワークは RBM (Restricted Boltzmann Machine)で構成し，教師なし学習を行う．教師なし学習が終わった RBM の出力側の層と出力との間の教師つき学習は，ロジスティック回帰を使ったものや，サポートベクターマシンを用いたもの，ニューラルネットワークを用いたものなど複数の構成が考えられるが，バックプロパゲーションを用いてファインチューニングすることによりネットワークを最適化する．5 節の実験では，ロジスティック回帰を用いたものを採用した．

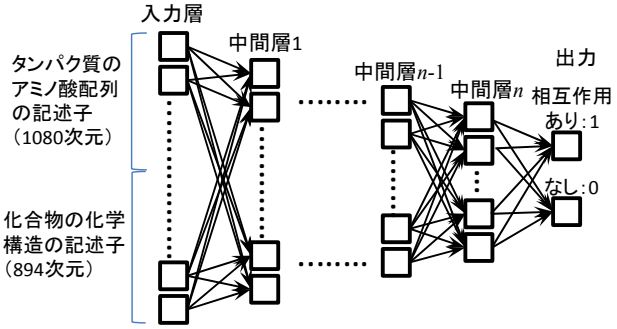

図 2：Deep Belief Networks

我々は，このように構築されたネットワークの重みを調べていくことで，今後，重要な特徴量を見つけることを目指している．

5. 実験結果

文献[4]では 1 万件の相互作用データを用いて適切なネットワーク構成を検討したが，5 万件のデータを用いて再度適切なネットワーク構成について検討する．次に，プレトレーニングの効果についての検証実験を行い，最後に，CGBVS 法と CGBVS-DBN 法の性能比較を行う．なお，DBN の構築は，Theano ライブラリ[13]を使用した Hinton らのコードに倣って作成した[7,8].

なお，DBN の評価ではデータの 4/6 を学習用データ 1/6 を評価用データ，1/6 をバリデーション用データとした．一方，SVM では，5-fold cross-validation で評価を行い，また，負例セットの違いによって性能が変化しないことを確認するため，負例セットを変化させた 5 セットのデータ作成し，それぞれのデータで学習した SVM の性能の平均で評価した．

DBN のネットワーク構成の検討

すべての中間層のユニット数が同じ構成に限定して適切な中間層数・ユニット数を検討する．中間層数は 1 から 9 まで 9 通り，ユニット数は，1000，2000，3000 の 3 通り，合計 27 通りについて性能を比較する．

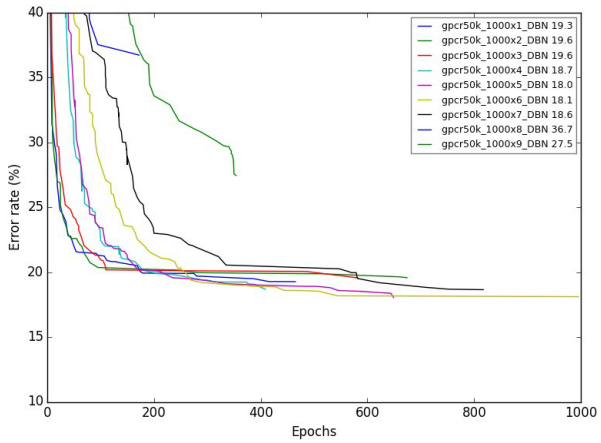

図 3：ユニット数 1000 で層数を変化させた結果

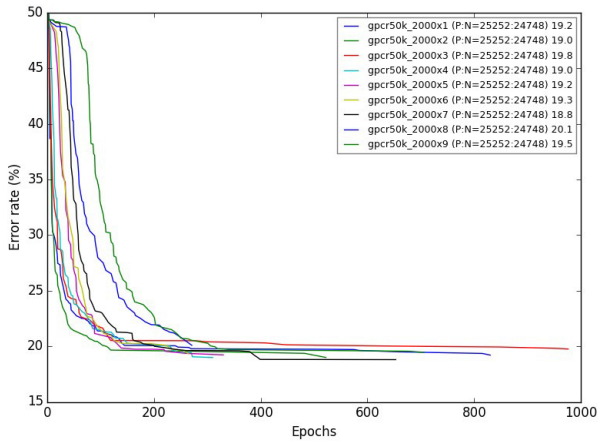

図 4：ユニット数 2000 で層数を変化させた結果

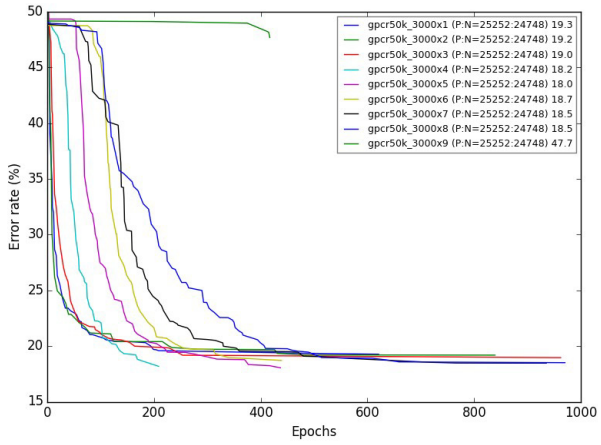

図 5：ユニット数 3000 で層数を変化させた結果

図 3 はユニット数が 1000 の場合，図 4 は 2000 の場合，図 5 は 3000 の場合での結果である．横軸はファインチューニングにおける Epoch 数で，縦軸はエラー率，すなわち，下になるほど性能が高い．

ユニット数 1000 の場合では，層数が 1 と 2 の場合のときに，他と比べて著しく性能が低くなった．これは，ネットワークの複雑さ（表現力）が低いために十分な学習ができなかったためである．

一方，ユニット数 3000 の場合では，9 層の場合で著しく性能が低くなった．これは，複雑なネットワークを学習するのに十分な量の学習データが与えられていなかったためである．

他の多くの場合では，エラー率は 18%から 20%の範囲に収まっていた．

以上より，一定以上の複雑さを持ったネットワークを用いれば，十分高い性能を発揮できることが確認された．また，ネットワークが単純すぎる場合や，複雑すぎる場合には，性能が低下することが確認された．最も性能が高かったのは，ユニット数 3000，層数が 5 のネットワークであった．

プレトレーニングあり/なしの性能比較

DBN は，プレトレーニングを行うことで，深い層のネットワークの学習を可能にしていると言われている．そこで，プレトレーニングを行った場合と行わなかった場合との性能を比較した．ネットワークは 3 層で各層いずれも 2000 ユニットの構成とした．

図 6 はその結果である．グラフ上の 2 本の曲線のうち，下側（緑）がプレトレーニングなしでファインチューニングを開始したもの，上側（青）がプレトレーニングを 100 回施した後にファインチューニングを開始したものである．エラー率が更新（低下）するごとにプロットを行っているため，プレトレーニングありは，Epoch 数 300 以降は性能が向上していない．

上記の結果から，プレトレーニングをした場合のほうが適切に学習され性能が向上するという予想に反し，今回のデータについてはプレトレーニングをしない場合のほうが性能が向上することを確認した．

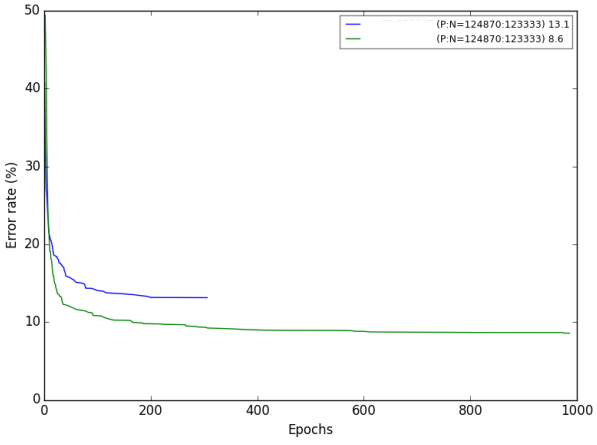

図 6：プレトレーニングあり/なしの比較

CGBVS 法と CGBVS-DBN 法の性能比較

CGBVS 法と CGBVS-DBN 法の性能比較を行った. CGBVS 法での SVM のハイパーパラメータの値は，文献[10]で述べられている値を用いた.

CGBVS-DBN 法では, DBN の構成は 5 層で各層いずれも 3000 ユニットのものを用い，プレトレーニングを行わずにファインチューニングを行った.

実験の結果，CGBVS 法の Accuracy が 91.4 であったのに対し, CGBVS-DBN 法では 91.7 となり，0.3 パーセントの性能向上であった.

表 1 : CGBVS 法と CGBVS-DBN 法の性能比較

	CGBVS-DBN 法	CGBVS 法
Accuracy	91.7%	91.4%

6. おわりに

本稿では, DBN に基づく化合物とタンパク質の相互作用を予測する手法を提案し, 25 万件の相互作用データを用いて，性能確認を行った．その結果，ある一定以上の複雑さ（表現力）を持ったネットワークであれば，Epoch 数を重なることにより性能に大きな違いがでないこと，プレトレーニングをしないほうが性能が高くなることを確認した.

さらに, SVM に基づく CGBVS 法と DBN に基づく CGBVS-DBN 法の性能を比較したところ，CGBVS-DBN 法のほうが高い性能を示すことを確認した.

今後，CGBVS-DBN 法を用いて，100 万件を超える相互作用データの学習を行っていく．また，大規模なデータを用いた場合での計算速度についても CGBVS 法と CGBVS-DBN 法でベンチマークを行っていきたい． さらに，構築した DBN を分析することで，化合物あるいはタンパク質のどの特徴が重要であるか解明していきたい.

謝辞

本研究の一部は，科学技術振興機構(JST)の戦略的創造研究推進事業(CREST)の助成を受けた.

参考文献

[1] Fujitani,H., Tanida, Y. and Matsuura, A.: Massively parallel computation of absolute binding free energy with well-equilibrated states, Phys. Rev. E, 79, 021914, 2009.

[2] Ewing, T. and Kuntz, I.: Critical Evaluation of Search Algorithms for Automated Molecular Docking and Database Screening, Journal of Computational Chemistry, 18:9, pp. 1175-1189, 1998.

[3] UCSF DOCK http://dock.compbio.ucsf.edu/.

[4] Morris, G. M., Goodsell, D. S., Halliday, R. S., Huey, R. Hart, W. E., Belew, R. K. and Olson, A. J.: Automated Docking Using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function, Journal of Computational Chemistry, 19:14, pp. 1639-1662, 1998.

[5] Jones. G., Willett, P., Glen, R. C., Leach, A. R., and Taylor, R.: Development and Validation of a Genetic Algorithm for Flexible Docking, Journal of Molecular Biology, 267:3, pp. 727-748, 1997.

[6] Yabuuchi, H., Niijima, S., Takematsu, H., Ida, T., Hirokawa, T., Hara, T., Ogawa, T., Minowa, Y., Tsujimoto, G. and Okuno, Y.: Analysis of multiple compound-protein interactions reveals novel bioactive molecules, Molcular Systems Biology, 7, p. 472, 2011.

[7] Hinton, G. E., Osindero, S. and The Y. W.: A fast learning algorithm for deep belief nets, Neural computation, Vol. 18, No. 7, pp. 1527-1554, 2006.

[8] Hinton, G. E., and Salakhutdinov, R. R.: Reducing the Dimensionality of Data with Neural Networks, Science, 313:5786, pp. 504-507, 2006.

[9] 浜中雅俊, 種石 慶, Brown, J. B., 奥野 恭史: 深層学習に基づくタンパク質と化合物の相互作用予測, 情報処理学会全国大会, 4B-07, 2015.

[10] Okuno, Y., Tamon, A., Yabuuchi, H., Niijima, S., Minowa, Y., Tonomura, K., Kunimoto, R. and Feng, C.: GLIDA: GPCR- Ligand Database for Chemical Genomics Drug Discovery - Database and Tools Update, Nucleic Acids Research, 36, D907-12, 2008.

[11] Rao, H.B., Zhu, F., Yang, G. B., Li, Z. R. and Chen, Y. R.: Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence, Journal of Nucleic Acids Research, 34:2, pp W32-W37, 2006.

[12] Andrea, M., Viviana, C., Manuela, P., Roberto, T.: DRAGON software: an easy approach to molecular descriptor calculations, MATCH Commun. Math. Computer Chem. 56:2, pp. 237-248, 2006.

[13] Bastien, F., Lamblin, P., Pascanu, R., Bergstra, J., Goodfellow, I., Bergeron, A. Bouchard, N., Warde-Farley, D. and Bengio, F.: Theano: new features and speed improvements, NIPS 2012 deep learning workshop, 2012.